**ANL 488 PROJECT PROPOSAL**

**Using Machine Learning to uncover insights**

**related to safety incidents (Mining Industry)**



**Submitted by Connie Tan @ Zann Tan**

**SCHOOL OF BUSINESS**

**Singapore University of Social Sciences**

**Presented to Singapore University of Social**

**Sciences in partial fulfilment of the**

**requirements for the**

**Degree of Bachelor of Science**

**in Business Analytics**

**2023**

Table of Contents

[Chapter 1: Business Understanding 2](#_Toc143071476)

[Chapter 1.1: Introduction 2](#_Toc143071477)

[Chapter 1.2: Business Problem & Objective 3](#_Toc143071480)

[Chapter 1.3: Data Mining Goal 4](#_Toc143071481)

[Chapter 2: Literature Review 5](#_Toc143071482)

[Chapter 3: Data Understanding & Preparation 10](#_Toc143071483)

[Chapter 3.1: Data Exploration 10](#_Toc143071484)

[Chapter 3.1.1: New Data 14](#_Toc143071488)

[Chapter 3.2: Data Preparation 14](#_Toc143071485)

[Chapter 3.2.1: Data Cleaning 15](#_Toc143071486)

[Chapter 3.2.2: Data Selection 17](#_Toc143071487)

[Chapter 3.2.4: Format Data 17](#_Toc143071489)

[Chapter 3.3: Data Specification 18](#_Toc143071490)

[Chapter 4: Proposed Modelling and Evaluation 19](#_Toc143071491)

[Chapter 5: Proposed Schedule 20](#_Toc143071492)

[References 21](#_Toc143071493)

# **Chapter 1: Business Understanding**

## Chapter 1.1: Introduction

|  |  |
| --- | --- |
| Mining is the method used to extract minerals from Earth, supplying the world with essential resources and creating employment opportunities. For example, coal mining created 6.5 million jobs worldwide (World Coal Association, 2023). Although mining offers employment advantages, miners often work in harsh work environments. Safe Work Australia (2022) recorded 2.3 fatalities per 100,000 workers and 130,195 total serious claims resulting in a median compensation of $15,072 per claim in the mining industry during 2021. Despite the hazardous environment in the mining sector, Manjunatha (2023) discovered only two out of 109 papers presented findings on occupational safety research suggesting safety is not widely studied. Alkaissy et al. (2023) cited insufficient injury data due to under-reporting, different recording methods and definitions (Safe Work Australia, 2013) including poor data quality were reasons for the lack of safety research.  *Figure 1: Location of mining sites in Australia and Canada* | |
|  |  |

For this study, datasets from a mining organization operating in Australia and Canada were used, illustrated in Figure 1. Additionally, Australian, and Canadian weather data were sourced from Kaggle to incorporate into the study.

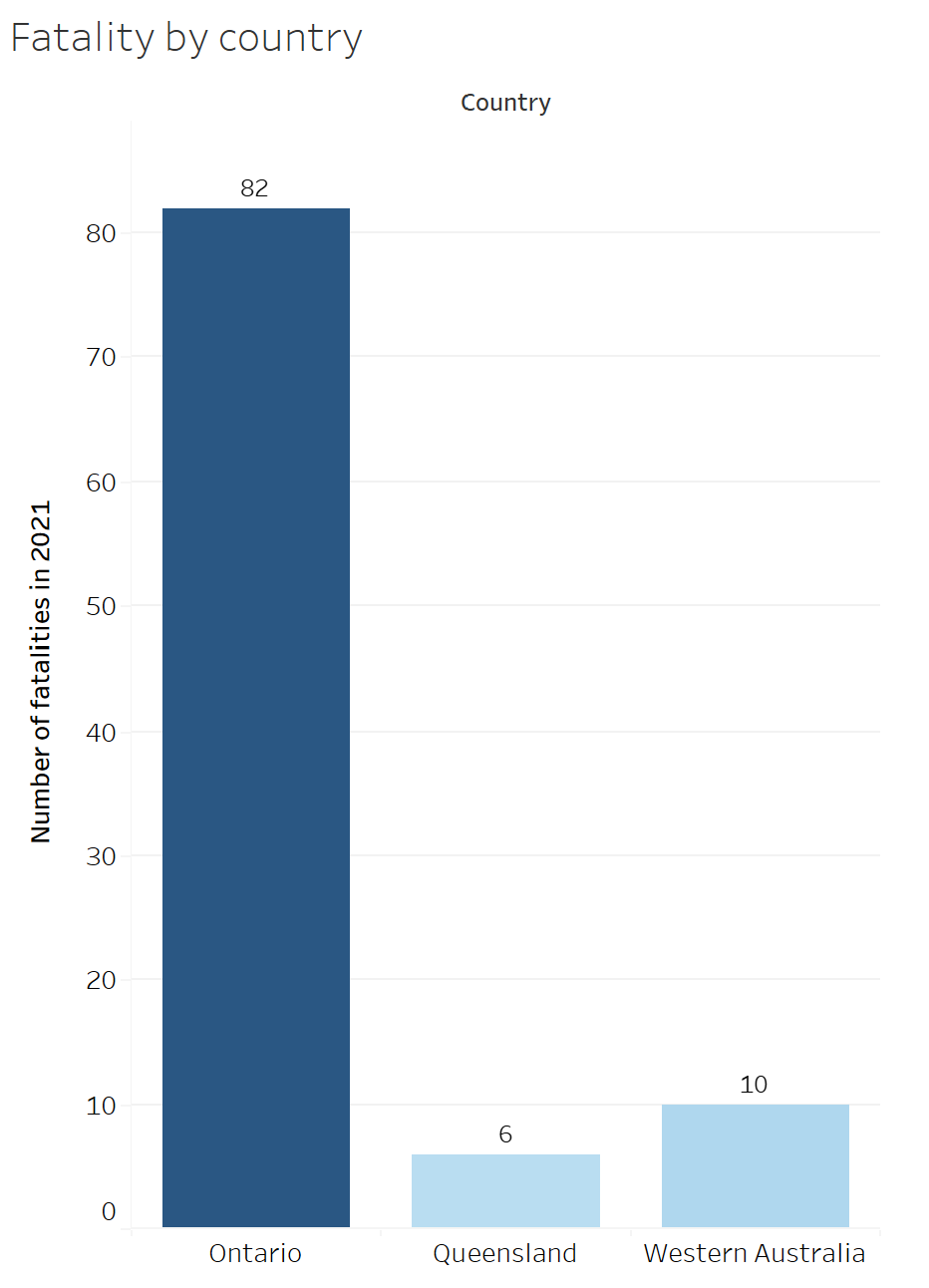
With insufficient safety knowledge from research, miners are constantly at risk of experiencing a safety incident, leading to detrimental effects on physical and mental well-being. Hence, this paper deploys machine learning alongside the CRISP-DM framework to investigate why safety incidents result in an injury or fatality to enhance safety performances in the mining industry through effective safety controls.

## Chapter 1.2: Business Problem & Objective

Miners are constantly exposed to different hazards at work ranging from chemical, ergonomic, health, physical, psychosocial, and safety (Government of Canada, 2023). For example, gas hazards, musculoskeletal disorders, lung diseases, heat stress, fatigue, confined spaces, and work at height (Queensland Government, 2023). To mitigate the risk of exposure to hazards, the government enacts legislation to protect miners. For example, in Canada, health and safety is governed by the “Occupational Health and Safety Act” in Ontario (Ontario’s Regulatory Registry, 2021). In Australia, health and safety is governed by the “Mining and Quarrying Safety and Health Act 1999” in Queensland (Resources Safety & Health Queensland, 2023) and the “Mines Safety and Inspection Act 1994” in Western Australia (Government of Western Australia, 2022).

Despite legislation, audits through inspection checks, best practices, and guidelines in Australia and Canada, mine operators struggle to achieve zero fatalities and injuries in the workplace. For example, Figure 2 illustrates fatality rates in Australia and Canada in 2021. Notably, Ontario, Queensland and Western Australia reported 82 (Workplace Safety and Insurance Board, 2023), 6 (Safe Work Australia, 2022), and 10 (*Minerals Safety Statistics*, 2023) fatalities respectively. Failing to achieve zero fatality rates highlights the importance of predictive safety being an iterative process to improve safety for miners to ensure a constant supply of minerals. Hence, the business problem centres around addressing fatalities and injuries in mines resulting in lost-time injuries (LTI). The business objective is to enhance the safety of miners in Australia and Canada by striving to attain zero fatalities and injuries at work.

*Figure 2: Fatality rates in Australia and Canada in 2021*



## Chapter 1.3: Data Mining Goal

The data mining goal is to predict factors likely to cause fatalities and injuries in mines using the given dataset. With new insights identified through machine learning beforehand, mining companies can take proactive actions to prevent fatalities and injuries, thereby, achieving zero fatalities and injuries, and keeping the workplace safe for everyone.

## Chapter 1.4: Software

This study utilizes four software to facilitate the research. Mainly, Excel and IBM SPSS for initial data exploration, Tableau for illustration, Python for data preparation, and IBM SPSS for initial proposed modelling and evaluation to assess the suitability of datasets for analysis.

# **Chapter 2: Literature Review**

Due to the complexity of different hazards present, human factors, tasks, and parties involved in safety incidents, Lee et al. (2020) employed a three-step framework to refine variables in the final dataset for modelling incident prediction.

Firstly, two methods, latent class cluster analysis (LCCA) and the chi-square test were applied for comparison. LCCA, a clustering technique, suitable for all data types (categorical, numeric, and binary) was applied to determine the optimum number of variables through an iterative process of experimenting with different values of “K” (latent class) by assessing the goodness of fit like Akiake information criteria (AIC)”, “Bayesian information criteria (BIC)”, “consistent AIC (CAIC)” and “R-squared” (Lee et al., 2020). Thereafter, Lee et al. (2020) applied the chi-square test to validate the findings presented by LCCA. For example, using LCCA, 130 variables in the original dataset were reduced to 7 variables in the final dataset, signifying a reduction of 94% of the variables from the original dataset. In this paper, a total of 61 variables were documented within the seven datasets provided.

Secondly, Lee et al. (2020) applied a support vector machine (SVM) and decision trees (DT) with ensemble to evaluate the number of categories within each variable through predictive performance. The research concluded ten categories per variable was the optimum number to analyse safety datasets. Following Lee et al. (2020)’s findings, this paper adheres to a maximum of ten categories per variable. Type of Occurrence Classification System (TOOCS) established by the Australian Safety and Compensation Council (2004) is incorporated to reduce the number of categories in the following variables: “AgencyOfInjury”, “BodyPart”, “MechanismOfInjury” and “NatureOfInjury”. Table 1 describes a detailed explanation of the different classification systems used. Lastly, Lee et al. (2020) performed correlation analysis and principal component analysis to prepare the final dataset. As the final dataset used in this study contains categorical variables, Stemn and Krampah (2022) recommend using correspondence analysis (CA), a statistical method suitable for categorical data to determine the correlation.

*Table 1: Detailed explanation of TOOCS*

|  |  |
| --- | --- |
| TOOCS classification | Explanation |
| Agency of injury | Recognizes the primary cause (object, substance, or circumstances) associated with the safety incident and respective causes directly responsible for the injury or disease |
| Body Location | Precisely identifies the injured body part |
| Mechanism of injury | Pinpoint the situation leading to the injury or disease |
| Nature of injury | Determine the injury or disease experienced by the worker including mental illness |

*Table 2: Summary of factors influencing safety incidents from literature reviews*

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Variables/  Paper | Datasets provided | Stemn and Krampah (2022) *“Injury severity and influence factors in surface mines: A correspondence analysis”* | *Butani (1988)*  *“Relative risk analysis of injuries in coal mining by age and experience at present company”* | Asare-Doku et al. (2022) *“Mental health and workplace factors: comparison of the Ghanaian and Australian mining industry”* | Stemn and Benyarku (2023)  *“Mineworkers’ perspective of fatigue: A study of the Ghanaian mining industry”* | Muzaffar et al. (2013)  *“Factors associated with fatal mining injuries among contractors and operators”* | Dumrak et al. (2013)  *“Factors associated with the severity of construction accidents: The case of South Australia”* | Vlachos (2019)  *“INTERRELATION BETWEEN OCCUPATIONAL HEALTH & SAFETY LEADING AND LAGGING INDICATORS IN MINING INDUSTRY. AN EMPIRICAL STUDY”* |
| Working hours | ✓ | ✓ |  |  |  | ✓ |  |  |
| Days away from work | ✓ | ✓ |  |  |  |  |  |  |
| Time of accidents |  | ✓ |  |  |  |  |  |  |
| Work experience | ✓ | ✓ | ✓ |  |  | ✓ | ✓ |  |
| Mental health status |  |  |  | ✓ | ✓ |  |  |  |
| Age |  | ✓ | ✓ |  |  |  | ✓ |  |
| Employment Type | ✓ |  |  |  |  | ✓ |  |  |
| Mine type |  |  |  |  |  | ✓ |  |  |
| Gender |  |  |  |  |  |  | ✓ |  |
| Size of organization |  |  |  |  |  |  | ✓ |  |
| Project size |  |  |  |  |  |  | ✓ |  |
| Mechanism of incident | ✓ |  |  |  |  |  | ✓ |  |
| Injury location | ✓ |  |  |  |  |  | ✓ |  |
| Nature of injury | ✓ |  |  |  |  |  | ✓ |  |
| Native language |  |  |  |  |  |  | ✓ |  |
| Safety indicators |  |  |  |  |  |  |  | ✓ |

Table 2 provides a summary of factors influencing safety incidents from seven literature reviews. Also, comparisons with the provided datasets were conducted to determine if similar data were available for this study.

To begin, Stemn and Krampah (2022) investigated the association between injury severity and accident factors (who, what, where, and why) because most studies focus either on fatalities or injuries without understanding how different accidents result in different injury severity. The research revealed influential injury severity factors including work experience, days away from work, and timing of accidents. Namely, young, and less experienced miners and night shifts were hidden causes of safety incidents. Indeed, similar findings were also derived by Butani (1988). Furthermore, frequent occurrences of safety incidents were during morning shifts, specifically between 12 pm to 5.59 pm, the second 6 hours of a 12-hour shift. Hence, understanding the relationship between incident factors and different injury severity helps to formulate effective safety strategies to reduce safety incidents. Additionally, another research by Dumrak et al. (2013) emphasizes factors including demographics (age, gender, native languages spoken), work experience, and environments that influence injury severity. For example, the proportion of injuries increases with age and experience. New and young workers were more prone to suffer from severe accidents due to lack of experience while older and experienced workers were more prone to fatal accidents due to overconfidence. Moreover, workers who did not speak English were also at higher risk of critical injuries due to the inability to understand safety training and manuals. Besides, the study explains how the mechanism of injury and injury location could help to further explain the severity of an injury, leading to more effective use of tools, equipment and materials.

Next, Asare-Doku et al. (2022) highlight mineworkers’ mental and social well-being are important factors in mitigating the risk of safety incidents. In the research, poor mental health status was attributed to long work hours, labour-intensive tasks, and environmental factors, resulting in fatigue and poor judgement, leading to safety incidents. Indeed, Stemn and Benyarku (2023) support the research that fatigue is a contributing factor to increased safety incidents. Besides, Muzaffar et al. (2013) explored factors associated with fatal incidents across employment types and concluded work experience, working hours, and mine type (surface or underground) were crucial to determining fatalities. For instance, statistically significant associations were observed between fatalities and contractors, lesser work experience, and more than 8 hours of work at surface mines.

Finally, Vlachos's (2019) challenges the accuracy of safety indicators used by mining organisations. The research reveals most safety indicators used by mining organizations serve as lagging indicators measuring past events instead of leading indicators which measure potential safety issues or areas requiring attention. Leading and lagging indicators are equally important because lagging indicator helps to assess the effectiveness of leading indicators in measuring safety performances. In particular, maintenance data serves as a leading indicator, to predict when a machine malfunctions and issues "warning alarms" when servicing is required (Benson, 2023). Consequently, Towsey (2011) questions the relevance of parameters used by organizations to represent safety. An example is the common use of lost-time injury frequency rates (LTIFR) to evaluate safety performance. Nevertheless, absolute measures of fatalities are preferred because analysis could be skewed to believe safety has improved due to fatalities being a rare event. Likewise, Safe Work Australia (2013) argues the use of LTI correlates negatively with influencing work injury and illness factors, resulting in the usage of ineffective measure safety performances. To illustrate, LTI with less severe injuries are more frequent than LTI with more severe injuries, potentially leading to a misleading impression of improved safety. Despite the challenges associated with LTI, organizations continue to adopt LTIFR to determine safety performance. Hence, Safe Work Australia (2013) recommends a severity framework that comprises injuries based on the mineworker’s impact.

*Figure 3: Injury severity framework*

Especially in this study, the framework to determine the severity of injury consists of lost time and injury classification as seen in Figure 3. Lost time is categorized as either “True” or ”False” and injury classification comprising of “First Aid Injury”, “Lost Time Injury”, “Medical Treatment Injury”, “No Treatment”, “Non-Work”, “Occupational Injury/Illness” and “Restricted Work Injury’. Figure 4 presents a breakdown of severity. It is noted that “Critical”, “Severe” and “Serious” accounts for 2.21%, 4.82% and 6.53% respectively are considered rare events.

*Figure 4: Proportion of severity*

A screenshot of a computer

Description automatically generated

*Figure 5: Methodology created based on reviewed literature*

Overall, based on the reviewed literature, a combination of the methodology illustrated in Figure 5 and machine learning techniques explains why safety incidents result in an injury or fatality.

# **Chapter 3: Data Understanding & Preparation**

## Chapter 3.1: Data Understanding

*Table 3: Summary of given datasets*

|  |  |  |  |
| --- | --- | --- | --- |
| Filename | Description of dataset | Rows | Columns |
| safety\_events.csv | Significant incident and reportable safety incidents from 2000 to 2022 | 6970 | 35 |
| production\_data.csv | Daily production data (tonnes) for each mine site from 2016 to 2023 | 7047 | 6 |
| emp\_start\_end\_dates.csv | Employment history from 2016 to 2022 | 1,048,575 | 3 |
| person\_workgroup.csv | Workgroup an employee belongs to during employment from 2020 to 2022 | 28856 | 4 |
| employee\_roster.csv | Roster and leave records of employees from 2019 to 2023 | 1,048,575 | 6 |
| labour\_hours\_worked.csv | Monthly labour hours worked by contractor/staff from 2011 to 2022 | 2078 | 4 |
| site\_location.csv | Location of mine sites | 10 | 3 |
|  | Total |  | 61 |

Seven datasets were provided, comprising 61 columns with more than 2 million rows of data, ranging from years between 2000 to 2023, illustrated in Table 3. During preliminary checks, data exploration was conducted using Excel and IBM SPSS, to identify any irrelevant columns, invalid data, and records of missing or duplicate values. To illustrate, invalid columns represent columns with duplicate information like “LTIDays” and “LTI” which conveyed similar information in different formats, numeric and flag respectively. Also, invalid data contains an invalid date format like “2999-01-01T00:00:00Z”. Table 4 presents a detailed observation of data quality issues during initial data exploration. Datasets are also combined to gather more information about employees. For example, invalid start dates such as “1990-01-01” in “person\_workgroup.csv” are replaced with “start\_dates” from “emp\_start\_end\_dates.csv” using “name\_hash” to determine the employee’s actual start date in the workgroup.

*Table 4: Detailed observation of data quality issues during initial data exploration*

|  |  |
| --- | --- |
| Dataset | Observation |
| person\_workgroup.csv | Using Excel, the following observations were made:  **Invalid date format**: from\_date, to\_date |
| employee\_roster.csv | No data quality issues were observed |
| labour\_hours\_worked.csv | No data quality issues were observed using Excel, but outliers were detected using IBM SPSS data audit node. |
| site\_location.csv | site\_key\_hashed in safety\_events.csv revealed missing 5 locations when mapped using Excel VLOOKUP with site\_location.csv. |
| emp\_start\_end\_dates.csv | Using Excel, the following observations were made:   1. **Invalid date format**: start\_date, end\_date 2. **NA values:** start\_date, end\_date 3. **Duplicate values**: 334 duplicate records in name\_hash (each row represents a specific employee, hence, should be unique) |
| safety\_events.csv | Using Excel, the following observations were made:   1. **Irrelevant columns:** day, event\_time, event\_reported\_time, EventId, AgencyOfInjuryId, BodyPartId, InjuryTypeCode,   MechanismOfInjuryId, shift\_commenced\_day,  shift\_commenced\_time, shift\_end\_day, shift\_end\_time,  AgencyOfInjuryDescription, BodyPartDescription,  NatureOfInjuryDecription, MechanismOfInjuryDecription   1. **Invalid date format:** event\_dt, event\_reported\_dt,   derived\_shift\_start\_dt, derived\_shift\_end\_dt   1. **NA values:** PersonName\_hashed, StaffContractor,   OrganisationName\_hashed, TimeBand  Using IBM SPSS data audit node, the following findings were revealed:   1. Absence of missing values except for outliers for LTIDays 2. Disproportionate proportion of “TRUE” and “FALSE” values for “Reportable” and “Significant” as seen below:   Using IBM SPSS distribution plot,   1. More than 10 categories were observed in AgencyOfInjury, BodyPart, Injury, MechanismOfInjury, NatureOfInjury |
| production\_data.csv | Using Excel, the following observations were made:   1. **Invalid date format in respective columns**: date, acutal\_tonnes\_moved,budgeted\_tonnes\_moved,short\_range\_forecast\_tonnes\_moved, half\_2\_forecast\_tonnes\_moved contains values with different decimal places.   Using IBM SPSS data audit node, the following findings were revealed:   1. No missing values except for outliers for   acutal\_tonnes\_moved,budgeted\_tonnes\_moved,short\_range\_forecast\_tonnes\_moved, half\_2\_forecast\_tonnes\_moved |

Lastly, an illustration in Figure 6 aids in comprehending the relationship among the seven datasets provided. For example, merging “safety\_events.csv” and “site\_location.csv” using “site\_key\_hashed” provides the location of the mining site.

*Figure 6: An illustration of how the seven datasets are inter-linked*

A screenshot of a computer screen

Description automatically generated

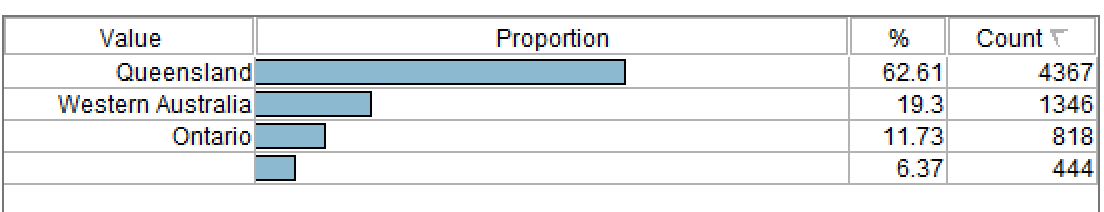
### Chapter 3.1.1: New Data

Using Kaggle, two new datasets on Australian weather data from Young & Young (2020), Canadian weather data from Turner (2020) and coordinates of mining locations in Ontario from the Ontario Ministry of Mines (2023) were incorporated to assess the impact of weather conditions on safety incidents. Firstly, the weather data retrieved includes daily records of rainfall and temperature alongside the respective cities in Australia and Canada. To illustrate in Figure 7, a methodical approach was used to select the nearest weather station in Perth, Brisbane, Townsville, and Ottawa to provide a better representation of the weather conditions experienced by mineworkers.

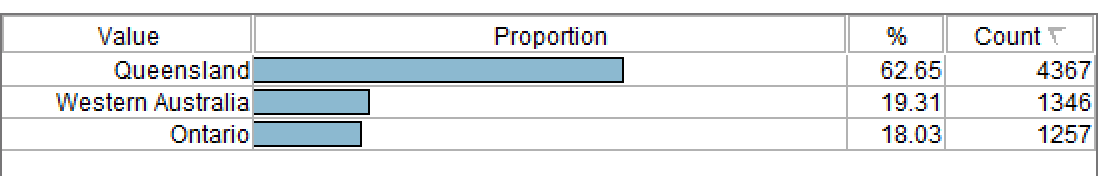
|  |  |
| --- | --- |
| *Figure 7: Location of mining sites in Australia and Canada* | |
|  |  |

Secondly, weather data obtained from Kaggle was merged into “site\_location.csv” through Python, with the unique key: “Region”. Subsequently, “site\_location.csv” was merged again with “safety\_events.csv” using “site\_key\_hashed” to derive an accurate representation of the actual weather conditions during the day of the incident. Lastly, coordinates of additional mining locations in “site\_location.csv” are merged using “site\_key\_hashed”. From Figure 8, 6.37% of missing values were replaced with Ontario mining location to balance the final dataset because Ontario contains only 11% of records, while Western Australia and Queensland contain 19.3% and 62.62% of records respectively. After merging, Ontario’s records contain 18.03% illustrated in Figure 9.

*Figure 8: Proportion of Ontario’s records (Before merging)*



*Figure 9: Proportion of Ontario’s records (After merging)*



## **Chapter 3.2: Data Preparation**

### Chapter 3.2.1: Data Cleaning

In this study, data cleaning adopts a five-step framework to ensure data quality.

Firstly, based on preliminary checks, irrelevant columns in Table 4 were removed.

Secondly, data cleaning was performed using Python on nine datasets including seven provided datasets and one weather dataset from Kaggle. The objective was to rectify invalid dates, eliminate duplicates, replace missing values, and remap the following variables: “AgencyOfInjury”, “BodyPart”, “Injury”, “MechanismOfInjury”, “NatureOfInjury”, “leave\_type” to a maximum of ten categories. For example, 36 categories within leave\_type from “employee\_roster.csv” were remapped to five categories.

*Table 5: Summary of new columns created after merging and percentage of N/A values*

|  |  |  |  |
| --- | --- | --- | --- |
| Files merged with safety\_events.csv | Unique keys | New column | Percentage of N/A values |
| production\_data.csv | site\_key\_hashed | ‘met\_target\_budgeted\_tones?, acutal\_tonnes\_moved, budgeted\_tonnes\_moved |  |
| labour\_hours\_worked.csv | site\_key\_hashed | ‘scheduled\_hours\_worked |  |
| ‘employee\_roster.csv | ‘name\_hash | Annual leave,  Compassionate Leave, Injury Leave, Maternity Leave, Medical Leave, Off Day, Parental Leave |  |
| ‘person\_workgroup.csv | ‘name\_hash | ‘switching\_roles, switching\_roles\_count |  |
| ‘employee\_roster.csv | ‘name\_hash | Total number of employees on leave |  |
| ‘emp\_start\_end\_dates.csv | ‘name\_hash | employee\_count,  employee\_count\_difference,  length\_of\_employment |  |

Thirdly, merging was executed using unique keys to incorporate human and organisational factors which could explain an occurrence of the safety incident. However, after merging the respective datasets, more missing (N/A) values were present. For instance, a merge using “name\_hash” between “employee\_roster.csv” and “person\_workgroup.csv” to understand the number of employees on leave within a workgroup during the time of the incident resulted in more missing values than the original dataset, depicted in Table 5. Subsequently, similar variables like “Annual leave” containing 99.78% of N/A values are non-informative for this study were removed to prepare the final dataset.

Fourthly, a series of new columns including “Incident\_time\_period”, “Same\_Date\_Reporting” and “Severity” was derived. For example, “Incident\_time\_period” is derived from the time of the incident to determine if there were significant findings relating to safety incidents. Also, “Severity” was calculated from a combination of “LostTime” and “Injury”, presented in Table 6.

*Table 6: Classification of severity*

|  |  |  |
| --- | --- | --- |
| LostTime | Injury | Severity |
| False | * Non work * No Treatment | Minor |
| False | * Occupational Injury/illness * First Aid Injury | Moderate |
| True | * Medical Treatment Injury | Serious |
| True | * Restricted Work Injury | Severe |
| True | * Lost Time Injury | Critical |

Lastly, LCCA, chi-square test and CA (to be discussed in Chapter 3.2.2) are applied to the cleaned dataset to select only variables exhibiting significant correlation for modelling.

### Chapter 3.2.2: Data Selection

In this study, the chi-square test and multiple correspondence analysis (MCA) are conducted to ascertain the relevance of the selected variables to be used for modelling. The chi-square test is a statistical technique to assess if any significant association exists between categorical variables (Hayes, 2023). Similarly, MCA is a multivariate statistical technique to identify associations among multiple categorical variables (Bock, 2022). Two statistical tests are deployed on the cleaned dataset comprising 13 variables.

Chi-square test

The chi-square test is applied to the independent variables (X) against the dependent variable (Y), “Severity” to assess if there were any significant associations between the variables. From Table 7, all independent variables are significantly associated with “Severity” except for “Daily\_Rainfall\_mm”.

*Table 7: Summary of chi-square test among independent and dependent variables*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Independent variable (X) | Dependent Variable (Y) | P-Value  (alpha = 0.05) | Significantly associated? | Screenshot of output from Python |
| AgencyOfInjury | Severity | 1.859296e-164 | Yes |  |
| BodyPart | Severity | 1.105019e-163 | Yes |  |
| Reportable | Severity | 0.0 | Yes |  |
| Significant | Severity | 1.181232e-42 | Yes |  |
| MechanismOfInjury | Severity | 6.008479e-140 | Yes |  |
| NatureOfInjury | Severity | 5.943718e-125 | Yes |  |
| Region | Severity | 2.802996e-201 | Yes |  |
| Weather\_Station | Severity | 2.802996e-201 | Yes |  |
| Same\_Date\_Reporting | Severity | 9.518596e-125 | Yes |  |
| Incident\_time\_period | Severity | 0.000007 | Yes |  |
| Daily\_Mean\_Temp\_degrees | Severity | 7.581276e-12 | Yes |  |
| Daily\_Rainfall\_mm | Severity | 0.394964 | No |  |

Multiple correspondence analysis

Correspondence analysis presents a unique perspective of categorical variables in the form of a two-dimensional biplot compared to traditional visualisations and focuses on relative relationships and associations among categorical data rather than presenting rows or columns with the highest record count by looking into the underlying patterns and relationships within categorical datasets (Bock, 2022). To interpret an MCA biplot, Bock (2022) suggests the following points:

1. “The further things are from the origin, the more discriminating they are”
2. “The closer things are to the origin, the less distinct they probably are”

For example, in Figure 10, “Severity\_Serious” and “Severity\_Severe” are further away “Daily\_Rainfall\_mm\_no\_rain”, “Daily\_Rainfall\_mm\_light”, “Severity\_Moderate” and “Severity\_Minor” suggesting an association between “Daily\_Rainfall\_mm\_no\_rain”, “Daily\_Rainfall\_mm\_light”, “Severity\_Moderate” and “Severity\_Minor” but no association between “Daily\_Rainfall\_mm\_light”, “Daily\_Rainfall\_mm\_no\_rain”, “Severity\_Serious”, “Severity\_Severe’ based on the dataset. This finding is supported by the chi-square test where “Daily\_Rainfall\_mm” is not significantly associated. Hence, “Daily\_Rainfall\_mm” will be removed from the final dataset.

From Figure 11, Firstly, “AgencyOfInjury\_Materials and Substances” is strongly associated with “Severity\_Moderate”. Secondly, “Severity\_Severe” is strongly associated with “AgencyOfInjury\_Environmental Agencies”, “AgencyOfInjury\_Chemicals and Chemical products” and “AgencyOfInjury\_Mobile plant and transport”. Next, “Severity\_Serious” is associated with “AgencyOfInjury\_Machinery”, “AgencyOfInjury\_Non-powered handtools”, and “AgencyOfInjury\_Powered Equipment”. Lastly, “Severity\_Minor” is associated with “AgencyOfInjury\_Not yet assessed” and “AgencyOfInjury\_Other”. This finding is supported by the chi-square test that “AgencyOfInjury” is significantly associated with “Severity”.

|  |  |
| --- | --- |
| *Figure 10: CA biplot of Daily\_Rainfall\_mm and Severity* | *Figure 11: CA biplot of BodyPart and Severity* |
|  |  |

## Chapter 3.3: Data Specification

*Table 8: Summary of final dataset for modelling*

|  |  |  |
| --- | --- | --- |
| Variable Name | Type | Values |
| Reportable | Categorical | True  False |
| Significant | Categorical | True  False |
| BodyPart | Categorical | Upper Limbs  Neck  Multiple Locations  Trunk  Lower Limbs  Head  Not yet assessed  Unspecified locations |
| MechanismOfInjury | Categorical | Falls, Trips and Slips of a person  Stress (Physical/Mental)  Hitting Objects with a part of the body  Heat, Electricity and Other Environmental Factors  Not yet assessed  Others and Unspecified Mechanisms of Incident  Chemicals and Other Substances  Being hit by moving objects  Sound and Pressure |
| Severity | Categorical | Serious  Severe  Critical  Moderate  Minor |
| NatureOfInjury | Categorical | Traumatic Joint/Ligament and Muscle/Tendon Injury  Burn  Wounds, Lacerations, Amputations and Internal Organ Damage  Other Injuries  Not yet assessed  Musculoskeletal and Connective Tissue Diseases  Mental Diseases  Intracranial Injuries  Fractures  Injury to nerves and spinal cord |
| Region | Categorical | Queensland  Western Australia  Ontario |
| Weather\_Station | Categorical | Townsville  Perth  Ottawa |
| Same\_Date\_Reporting | Categorical | Yes  No |
| Incident\_time\_period | Categorical | Evening  Midnight  Morning  Afternoon |
| Daily\_Mean\_Temp\_degrees | Categorical | Cold  Mild |

Table 8 provides a summary of the final dataset including the variable name, type, and values.

# **Chapter 4: Proposed Modelling and Evaluation**

The proposed modelling would adopt supervised machine learning like decision trees, neural networks, support vector machines and random forests where results will be evaluated to determine the champion model. For instance, the target variable is “Severity” where the model will predict the severity of an injury based on the mechanism of injury, body injury, agency of injury, nature of injury, location, and weather data. With these insights, mining organisations can adopt safer control measures to reduce the occurrence of injuries. Figure 12 presents the proposed modelling of “Severity” using a neural network alongside C5.0 for better interpretation. Additionally, Figure 13 illustrates the model performance of 82.41% and 81.23% for training and testing respectively for neural networks. Moving forward, further tuning and models will be explored.

Figure 12: Proposed modelling in IBM SPSS

A diagram of a diagram

Description automatically generated

Figure 13: Model performance

A white sheet with black text

Description automatically generated

# **Chapter 5: Proposed Schedule**

The proposed schedule is presented in Table 9.

*Table 9: Proposed schedule*

|  |  |
| --- | --- |
| Dates | Task |
| 11 to 15 September | Review of literature reviews on machine learning approaches |
| 16 to 22 September | Machine learning modelling using Python |
| 23 Sept to 1 October | Document findings, and recommendations in the final report |
| 2 October to 6 October | Prepare for oral presentation |
| 7 October to 5 November | Finalize final report based on feedback from project proposal and oral presentation |

# **References**

Alkaissy, M., Arashpour, M., Golafshani, E. M., Hosseini, M. R., Khanmohammadi, S., Bai, Y., & Feng, H. (2023). Enhancing construction safety: Machine learning-based classification of injury types. *Safety Science*, *162*, 106102. https://doi.org/10.1016/j.ssci.2023.106102

Asare-Doku, W., Jane, R. L., Kelly, B., Amponsah-Tawiah, K., & James, C. (2022). Mental health and workplace factors: comparison of the Ghanaian and Australian mining industry. *BMC Health Services Research*, *22*(1). https://doi.org/10.1186/s12913-022-07712-0

Bock, T. (2022, April 25). *How to interpret correspondence analysis plots (It probably isn’t the way you think) - Displayr*. Displayr. https://www.displayr.com/interpret-correspondence-analysis-plots-probably-isnt-way-think/

Butani, S. (1988). Relative risk analysis of injuries in coal mining by age and experience at present company. *Journal of Occupational Accidents*, *10*(3), 209–216. https://doi.org/10.1016/0376-6349(88)90014-4

Dumrak, J., Mostafa, S., Kamardeen, I., & Rameezdeen, R. (2013). Factors associated with the severity of construction accidents: The case of South Australia. *Construction Economics and Building*, *13*(4), 32–49. https://doi.org/10.5130/ajceb.v13i4.3620

Government of Canada. (2023, July 13). *CCOHS: Hazards*. Retrieved August 12, 2023, from https://www.ccohs.ca/topics/hazards/#ctgt

Government of Western Australia. (2022, March 31). *WALW - Mines Safety and Inspection Act 1994 - home page*. Western Australian Legislation. Retrieved August 15, 2023, from https://www.legislation.wa.gov.au/legislation/statutes.nsf/main\_mrtitle\_599\_homepage.html

Hayes, A. (2023). Chi-Square (χ2) Statistic: What It Is, Examples, How and When to Use the Test. *Investopedia*. https://www.investopedia.com/terms/c/chi-square-statistic.asp

Lee, J., Yoon, Y., Oh, T. K., Park, S., & Ryu, S. I. (2020). A study on Data Pre-Processing and Accident Prediction Modelling for occupational accident analysis in the construction industry. *Applied Sciences*, *10*(21), 7949. https://doi.org/10.3390/app10217949

Manjunatha, A. (2023, May 12). *Injury Prediction in Mining Industry through Applied Machine Learning Approaches  - NORMA@NCI Library*. National College of Ireland. Retrieved August 15, 2023, from https://norma.ncirl.ie/6559/

*Minerals safety statistics*. (2023, February 20). Department of Mines, Industry Regulation and Safety. Retrieved August 15, 2023, from http://dmp.wa.gov.au/Safety/Safety-statistics-16198.aspx

Muzaffar, S., Cummings, K. J., Hobbs, G. R., Allison, P. D., & Kreiss, K. (2013). Factors associated with fatal mining injuries among contractors and operators. *Journal of Occupational and Environmental Medicine*, *55*(11), 1337–1344. https://doi.org/10.1097/jom.0b013e3182a2a5a2

Ontario Ministry of Mines. (2023). *Map*. Retrieved August 16, 2023, from https://oma.on.ca/en/ontario-mining/Map.aspx

Ontario’s Regulatory Registry. (2021, July 28). *Mining Health and Safety Regulatory Amendment Proposal*. © King’s Printer for Ontario, 2022. Retrieved August 15, 2023, from https://www.ontariocanada.com/registry/view.do?postingId=37907

Queensland Government. (2023, June 9). *Mining hazards database*. Business Queensland. Retrieved August 15, 2023, from https://www.business.qld.gov.au/industries/mining-energy-water/resources/safety-health/mining/hazards/hazards

Resources Safety & Health Queensland. (2023, July 4). *What we do*. Resources Safety and Health Queensland. Retrieved August 15, 2023, from https://www.rshq.qld.gov.au/about-us/what-we-do

Safe Work Australia. (2013). ISSUES IN THE MEASUREMENT AND  REPORTING OF WORK HEALTH AND  SAFETY PERFORMANCE: A REVIEW. In *Safeworkaustralia*. Retrieved September 5, 2023, from https://www.safeworkaustralia.gov.au/system/files/documents/1703/issues-measurement-reporting-whs-performance.pdf

Safe Work Australia. (2022, November 7). *Key work health and safety statistics Australia 2022*. Retrieved August 15, 2023, from https://www.safeworkaustralia.gov.au/doc/key-work-health-and-safety-statistics-australia-2022

Stemn, E., & Benyarku, C. A. (2023). Mineworkers’ perspective of fatigue: A study of the Ghanaian mining industry. *Safety Science*, *162*, 106095. https://doi.org/10.1016/j.ssci.2023.106095

Stemn, E., & Krampah, F. (2022). Injury severity and influence factors in surface mines: A correspondence analysis. *Safety Science*, *145*, 105495. https://doi.org/10.1016/j.ssci.2021.105495

Turner, A. (2020, January 27). *Eighty years of Canadian climate data*. Kaggle. Retrieved September 6, 2023, from https://www.kaggle.com/datasets/aturner374/eighty-years-of-canadian-climate-data?resource=download

Vlachos, T. (2019). INTERRELATION BETWEEN OCCUPATIONAL HEALTH & SAFETY LEADING AND LAGGING INDICATORS IN MINING INDUSTRY. AN. . . *ResearchGate*. https://www.researchgate.net/publication/338264721\_INTERRELATION\_BETWEEN\_OCCUPATIONAL\_HEALTH\_SAFETY\_LEADING\_AND\_LAGGING\_INDICATORS\_IN\_MINING\_INDUSTRY\_AN\_EMPIRICAL\_STUDY

Workplace Safety and Insurance Board. (2023, June 29). *Health and safety statistics*. Safetycheck. Retrieved August 15, 2023, from https://safetycheck.onlineservices.wsib.on.ca/safetycheck/explore/provincial/SH\_12/fatality

World Coal Association. (2023, March 20). *Coal’s contribution - World Coal Association*. Retrieved August 15, 2023, from https://www.worldcoal.org/coal-facts/coals-contribution/

Young, J., & Young, A. (2020, December 11). *Rain in Australia*. Kaggle. Retrieved September 6, 2023, from https://www.kaggle.com/datasets/jsphyg/weather-dataset-rattle-package